Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges

Year: 2021 | Citations: 829 | Authors: C. Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova

---

Abstract

Interpretability in machine learning (ML) is crucial for high stakes decisions and troubleshooting. In this work, we provide fundamental principles for interpretable ML, and dispel common misunderstandings that dilute the importance of this crucial topic. We also identify 10 technical challenge areas in interpretable machine learning and provide history and background on each problem. Some of these problems are classically important, and some are recent problems that have arisen in the last few years. These problems are: (1) Optimizing sparse logical models such as decision trees; (2) Optimization of scoring systems; (3) Placing constraints into generalized additive models to encourage sparsity and better interpretability; (4) Modern case-based reasoning, including neural networks and matching for causal inference; (5) Complete supervised disentanglement of neural networks; (6) Complete or even partial unsupervised disentanglement of neural networks; (7) Dimensionality reduction for data visualization; (8) Machine learning models that can incorporate physics and other generative or causal constraints; (9) Characterization of the"Rashomon set"of good models; and (10) Interpretable reinforcement learning. This survey is suitable as a starting point for statisticians and computer scientists interested in working in interpretable machine learning.